Journal of Economics and Financial Analysis
Type: Double Blind Peer Reviewed Scientific Journal
Printed ISSN: 2521-6627 | Online ISSN: 2521-6619
Publisher: Tripal Publishing House | DOI:10.1991/jefa.v1i1.a1
Journal homepage: www.ojs.tripaledu.com/jefa

Does Statistical Significance Help to Evaluate Predictive Performance of Competing Models?

Levent BULUT*
Department of Economics, University of Georgia, United States

Abstract

In Monte Carlo experiment with simulated data, we show that as a point forecast criterion, the Clark and West's (2006) unconditional test of mean squared prediction errors does not reflect the relative performance of a superior model over a relatively weaker one. The simulation results show that even though the mean squared prediction errors of a constructed superior model is far below a weaker alternative, the Clark- West test does not reflect this in their test statistics. Therefore, studies that use this statistic in testing the predictive accuracy of alternative exchange rate models, stock return predictability, inflation forecasting, and unemployment forecasting should not weight too much on the magnitude of the statistically significant Clark-West tests statistics.

Keywords: *Model comparison; Predictive accuracy; Point-forecast criterion; The Clark and West test; Monte-Carlo methods; Forecast comparison.*

JEL Classification: *F37, F47, G17, C52.*

* E-mail addresses: leventbulut@me.com

1. Introduction

In economics and finance, one commonly used approach for model selection and model comparison is called point-forecast criterion where out-of-sample forecasts computed from ex-post data are utilized to compare the mean squared prediction errors (MSPEs) across competing models. For that purpose, Diebold and Mariano (1995) and West (1996) proposed a test of equal predictability of two non-nested models. Moreover, Clark and West (2006) introduced the test of equal predictability for nested models. In this approach, a quadratic loss function is defined as the square of the prediction error. Then, a loss-differential function is calculated as the difference between the loss function of the null model and the structural model. The literature uses the point forecast criterion in testing the predictive accuracy of alternative exchange rate models [1], stock return predictability[2], inflation forecasting[3], and unemployment forecasting[4].

The way the literature compares different models is through comparison of the out-of-sample forecasting power of each contestant model against a null that is whether a driftless random walk or random walk with a drift. This practice comes with a drawback: with this testing procedure, the relative performance of each theory-driven contestant model against a second alternative cannot be evaluated. Instead, one can only measure the predictive accuracy against the null. How about the relative magnitude of the estimated test statistics? How much information we can capture about relative model performance from having a statistically significant test statistics for Model A that is, say, two times bigger than a statistically significant test statistic for Model B? In this paper, we answer these questions by generating two series that have nonlinear relations by construction and compare the in-sample fits and out-of-sample forecasts of linear and non-parametric models in light of the point forecast criterion. Then, we conduct Monte Carlo simulation to look at the finite sample properties of the Clark & West (CW test, henceforth) test statistics.

Our findings show that the CW test fails to reflect the relative performance of the non-parametric estimations over the OLS model. Even though the non-parametric model has the lowest MSPE, since we divide the mean loss differential function by sample estimate of the standard deviation of the loss-differential

[1] Some recent examples are Engel et.al (2015), Nikolsko-Rzhevskyy and Prodan (2012), Wang and Wu (2012), Ferraro et.al (2015) amongst many others.

[2] Some recent applications are Dimpfl and Jank (2015), Wu and Lee (2015), Löffler (2013) and Sousa et.al (2016).

[3] Faust and Wright (2013) and Arai (2014) are some examples in this area.

[4] Gregory and Zhu (2014) and Hutter and Weber (2015) are some recent examples.

function, the CW test produces lower values for the NP model relative to that of OLS model. Hence, comparing each model against the null does not give an accurate picture of the relative performance of one alternative to another one.

Section 2 summarizes the Diebold and Mariano (1995) and West (1996) and Clark and West (2006) approach to predictive accuracy testing. Section 3 introduces simulation practice where two series that have a non-linear relation by construction are created to look at the in-sample and out-of-sample performance of OLS and non-parametric model. Section 4 looks at the limiting distribution of the CW test using a Monte-Carlo experiment and Section 5 concludes.

2. Point Forecast Criterion

In the point forecast criterion, the out-of-sample forecast performance of contestant models is evaluated by comparing MSPEs. After defining a quadratic loss function $L(.)$ as the square of the forecast error, a loss differential function is defined as the difference between the loss function of the benchmark model and the structural model. We can define the loss function for the benchmark model b, $L(y_t^b)$, and the structural model sm, $L(y_t^{sm})$, as in (1):

$$L(y_t^b) = (y_t - \hat{y}_t^b)^2$$
$$L(y_t^{sm}) = (y_t - \hat{y}_t^{sm})^2 \quad (1)$$

where y_t is the actual series, $\hat{y}_t^{sm}$ and $\hat{y}_t^b$ are the forecasts obtained from the structural model sm and the benchmark model b, respectively. Then, the forecast accuracy testing is based on whether the population mean of the loss differential series d_t is zero where:

$$d_t = L(y_t^b) - L(y_t^{sm}) = (y_t - \hat{y}_t^b)^2 - (y_t - \hat{y}_t^{sm})^2 \quad (2)$$

In Diebold & Mariono (1995) and West (1996) (DMW test), under the null, the distribution of the sample mean of loss-differential $E[d_t]$ is asymptotically standard normal. However, Clark and West (2006) show that in cases where the alternative model nests the null model, DMW test is not suitable because for nested models since the alternative model has a large number of predictors than the null, it will produce a noise. Therefore, the sample MSPE difference is positive and an adjustment term is needed to center the DMW test statistics around zero. As for the adjustment, Clark and West (2006) suggest an adjusted MSPE for the alternative model that is adjusted downwards to have equal MSPEs under the null. Accordingly they propose the following adjusted loss-differential function:

$$d_t - adj_t = L(y_t^b) - \{L(y_t^{sm}) - adj_t\} \\ = (y_t - \hat{y}_t^b)^2 - (y_t - \hat{y}_t^{sm})^2 + (\hat{y}_t^b - \hat{y}_t^{sm})^2 \quad (3)$$

If $\tilde{d}$ indicates the mean of the adjusted-loss differential function in (3), $\widehat{avar}(\tilde{d})$ is the variance of $\tilde{d}$, then the CW test takes the following form:

$$CW = \frac{\tilde{d}}{(\widehat{avar}\,(\tilde{d}))^{1/2}} \quad (4)$$

In testing the predictive accuracy of alternative exchange rate models and stock return predictability, the literature uses driftless random walk or random walk with drift as the benchmark (null) model. In the case when the null model is driftless random walk, out-of-sample forecast for the null takes the value of zero ($\hat{y}_t^b = 0$) and the mean of the adjusted loss differential function $\tilde{d}$ in Clark and West (2006) in (4) takes the following form:

$$\tilde{d} = MSPE^{RW} - [MSPE^{sm} - P^{-1}\sum_{t=R+1}^{T}(\hat{y}_t^{sm})^2] \quad (5)$$

where P is the number of out-of-sample forecasts, R refers to rolling regression window size, the $MSPE^{RW}$ is the mean squared prediction error for the null, and the $MSPE^{sm}$ is the mean squared prediction error for the structural model. On the other hand, when the null is random walk with drift, $\tilde{d}$ in Clark and West (2006) in (4) takes the following form:

$$\tilde{d} = MSPE^{RW} - [MSPE^{sm} - P^{-1}\sum_{t=R+1}^{T}(\bar{y}_t - \hat{y}_t^{sm})^2] \quad (6)$$

where $\bar{y}_t$ is the estimate of draft at time *t* and in our paper, it is calculated as the average of *y* for observations 1 through *t*.

3. Simulation Practice

To better gauge the performance of the CW test statistics in (4), we implement simulation practice by artificially generating two series that have nonlinear relations by construction. For that purpose, we take the monthly actual real exchange rate[5] data for Australia and call it x. The variable x has a mean of 27.5 and standard deviation of 16.2 with a sample size of 399. Then, we generate time series data of *y* from *x* by using the following equation and call it the true model:

$$y = 4 + 0.5x + 0.10x^2 + e \quad (7)$$

[5] There is no specific reason to choose real exchange rate data; it could be any macro data.

where e is a random draw from a normal distribution with a mean of zero and standard deviation of ten. Given that the true model in (7) is a non-linear one, we fit the data to a linear model and a nonparametric model to see which one does a better job both in-sample and out-of-sample. Hence, we have the following two competing models where the latter one is the correct model:

$$\begin{aligned} &Model\ 1: (OLS) \quad y = \alpha + \beta x + \varepsilon \\ &Model\ 2: (NP) \quad y = g(x) + \varepsilon \end{aligned} \tag{8}$$

3.1. In-sample Comparison

First, we estimate the fitted values based on Model 1 and Model 2. In Model 1, we simply use the OLS method to get the predicted values. For Model 2, since the actual model is a non-linear one, we use the local linear kernel estimation as it has less bias compared to the local constant kernel estimation and show the results in Figure 1.

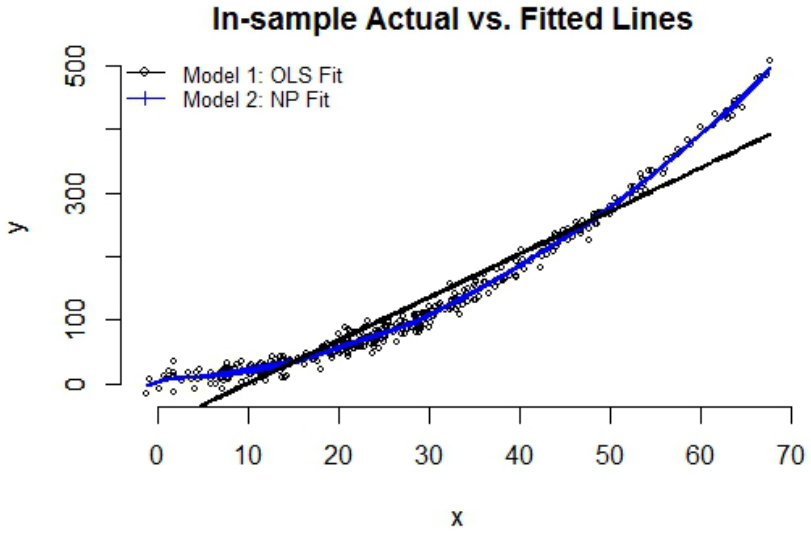

Figure 1. In-sample Actual vs. Fitted Lines

Figure 1 shows the (in-sample) real and fitted values for Model 1 (in black) and Model 2 (in blue). As expected, Model 2 outperforms Model 1 in in-sample fit. Accordingly, in-sample mean squared prediction errors (MSPEs) reflect the performance of each model. As expected, since Model 2 (NP Model) is correct, it

has low MSPE than the OLS model. The results show that OLS model has an in-sample MSPE of 946 while NP Model has 102.

3.2. Out-of-sample Comparison

We use the rolling regression method to get the one-month ahead out-of-sample forecasts. There are 399 observations in the sample. We choose R=200 as the window size and generate out-of-sample forecasts from these models. In out-of-sample forecasting, the OLS model (Model 1) produces MSPE of 560 while this number is just 117 for the NP model (Model 2). Hence, the results show that Model 2 does an excellent job in out-of-sample forecasting as well. Do these two models produce statistically significant different MSPEs at the population? Each model is tested against a benchmark model to answer this question. The benchmark model is whether driftless random walk or a random walk with a drift.

It is worth mentioning that the point forecast approach is heavily used in the literature on exchange rate and stock return predictability. However, the benchmark models (driftless random walk or random walk with drift) are for the log of exchange rate series and the log of stock market prices in *levels*. On the other hand, the literature mostly explains the exchange rate and stock price data in log-difference form (return series) due to stationarity of the data in levels. If p_t is the price index in logged (whether exchange rate or stock price), a driftless random walk model for the price series would be: $p_t = p_{t-1} + \varepsilon_t$ and random walk with a drift for the price series would be: $p_t = \alpha + p_{t-1} + \varepsilon_t$. Hence, if we define series in difference form, as shown below, the driftless random walk will be $y = p_t - p_{t-1} = \varepsilon_t$. For the random walk with a drift null, the model will be $y = p_t - p_{t-1} = \alpha + \varepsilon_t$.

Therefore, to compare models, we use the driftless random walk (Model 3) and random walk with a drift (Model 4) as two benchmarks (the null) and by using the CW test; we look at the out-of-sample forecast performances of Model 1 and Model 2.

$\boldsymbol{Model\ 1}: (OLS) \quad y = \alpha + \beta x + \varepsilon$ $\qquad$ $\boldsymbol{Model\ 2}: (NP) \quad y = g(x) + \varepsilon$

$\boldsymbol{Model\ 3}: (Driftless\ RW) \quad y_t = \varepsilon$ $\qquad$ $\boldsymbol{Model\ 4}: (RW\ with\ drift) \quad y_t = \alpha + \varepsilon$

Out-of-sample predictions for the driftless random walk (Model 3) for observations R+1, …… 399 are: $0,0,0 \dots \dots \dots \dots \dots \dots \dots \dots ,0$. Out-of-sample predictions for the random walk with a drift (Model 4) for observations R+1, ……,399 are: $\bar{y}_R, \bar{y}_{R+1}, \dots \dots \dots \dots \dots \dots \dots \dots , \bar{y}_{399}$, where $\bar{y}_R$ is the sample average for observations 1,…R. Figure 2 below shows the plots of out-of-sample forecast

errors for Models 1 and 2. As expected, the prediction errors for the NP model fluctuate around zero, while, for the OLS model, there are too many deviations.

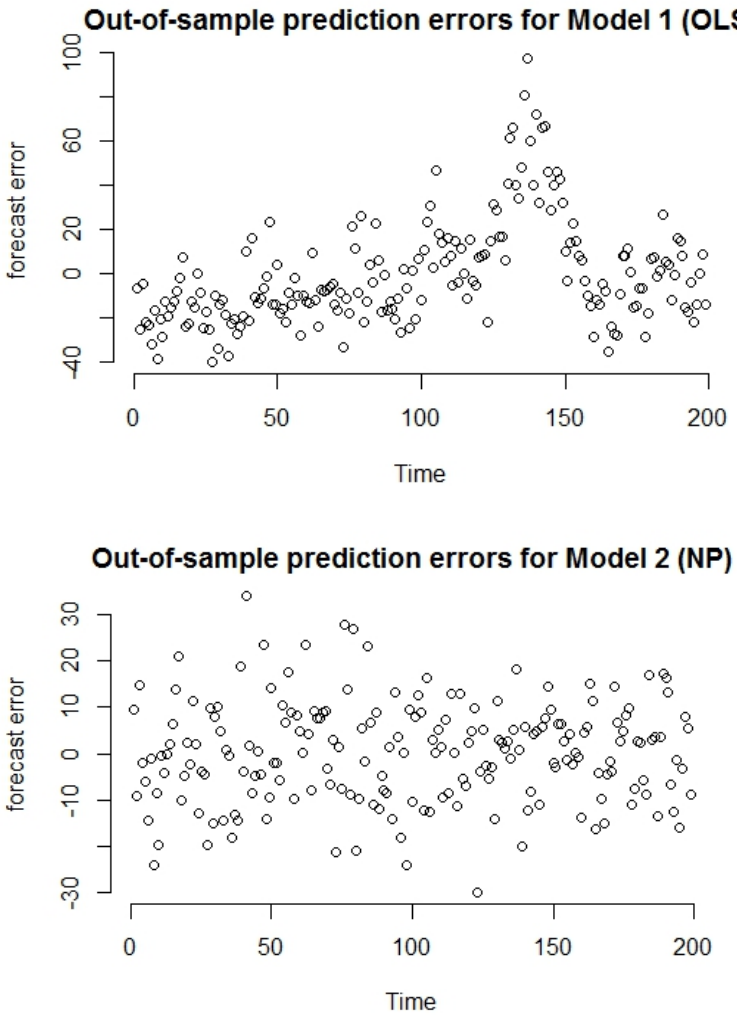

Figure 2. Out-of-sample Prediction Errors

Also, we look at the plot of NP forecast errors against the *x* variables and show it in Figure 3. The average out-of-sample forecast errors fluctuate around zero regardless of the level of *x* (except for a few observations). Therefore, the fit does not over smooth the data.

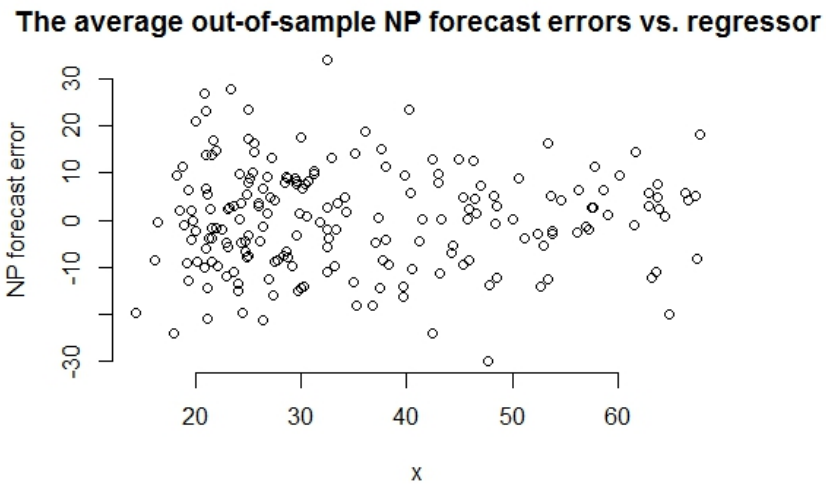

Figure 2. Residual Plots for NP Model

As for out-of-sample forecast performance of each model, we need to compare the sample MSPE of each contestant model against the null. Therefore, we use the CW test statistics for the predictive accuracy testing. We show the test results in Table 1. The null models produce huge MSPEs (MSPE for the dritfless random walk is 42,827 and is 14,065 for the random walk with a drift null). Therefore, we expect the CW test statistics to produce statistically significant positive values.

The results in Table 1 indicate that while MSPE for the NP estimation is far below the OLS, the magnitude of the CW test statistics fails to show the relative performance of Model 2 against Model 1. However, the nature of the test is such that we do not compare one alternative model with another one per se, but we compare each model with the null. Therefore, we need to examine how each model does a good job compared to the null (driftless random walk or the random walk with drift model). To better clarify this outcome, we look at the sample estimate of the variance of the loss-differential function $\tilde{d}$ in Table 2 below.

Table 1. Out-of-sample Forecast Comparisons with the Clark and West Test

Null: Driftless RW		**Null: RW with drift**	
Model 1	11.0900 (0.0000)	Model 1	10.8200 (0.0000)
Model 2	10.3000 (0.0000)	Model 2	9.6900 (0.0000)

Notes: The table shows the Clark and West (CW) test results for Model 1 (OLS) and Model 2 (NP). We use rolling regression method with a window of 200 observations. In the NP forecasts, we compute the least-squares cross-validated bandwidths for the local linear estimators. P-values are in the parantheses. The limiting distribution of the CW test under the null is standard normal. The statistically significant CW test indicate that the structural model outperforms the null.

Table 2 shows that, against the null of the driftless random walk, the variance of the adjusted loss-differential function for Model 2 is 28% higher than for Model 1. This difference reaches to 44% when the null is a random walk with drift. Even though NP has a far lower MSPE, since we divide the mean loss differential function by a larger number, the CW test produces lower values for the NP model relative to that of OLS model. Our findings suggest that comparing each model in reference to the null does not give an accurate picture of the relative performance of one alternative to another in reference to the correct data.

Table 2. Variance of the Adjusted Loss-differential Function

Null: Driftless RW		**Null: RW with drift**	
Model 1	10690.31	Model 1	4683.809
Model 2	13656.02	Model 2	6778.881

Notes: The table shows the variance of the adjusted loss differential function $\tilde{d}$ in Clark and West (CW) test under alternative null models.

We also look at the Theil's (1966) U value to check if we can find similar results. Theil's U value looks at the square root of the MSPE of the structural model over the square root of the MSPE of the random walk model as shown in (9):

$$TU = \frac{\sqrt{MSPE^{sm}}}{\sqrt{MSPE^{rw}}} \quad (9)$$

A TU value lower than one implies that the alternative model outperforms the null model[6]. In contrast to Clark and West test statistics, Theil's U test provides, shown in Table 3, consistent results regarding predictive accuracy. Model 2, which has a lower MSPE, has a smaller TU value than Model 1, indicating that it has a better out-of-sample forecast.

Table 3. The Theil's U-Statistics

Null: Driftless RW		**Null: RW with drift**	
Model 1	0.1100	Model 1	0.2000
Model 2	0.0500	Model 2	0.0900

Notes: The table shows the calculated Theil's U-statistics for Model 1 (OLS) and Model 2 (NP) under the null of driftless random walk and random walk with a drift. The higher the Theil's U-statistics, the lower the predictive performance.

4. Monte Carlo Experiment

We also perform Monte Carlo experiment to track the population distribution of MSPEs of structural models as well as the Clark and West test statistics. In fact, checking the performance of the Clark and West test in this analysis is trivial because of the nature of the simulation. Our null models, Model 3 and Model 4, perform poorly in out-of-sample. Therefore, CW test will always reject the null of equal MSPEs.

In examining how each model performs in repeated samples with Monte Carlo experiment, we take *x* series as given and use the same true parameters of the model in equation (7). Then, we generate new *y* series for a different e series drawn from the same distribution and repeat this exercise 500 times. Then, for each draw, we compute new out-of-sample forecasts, calculate the corresponding MSPEs and the CW test statistics, and present the results in Table 4.

[6] The Theil (1966)'s *U* statistic falls between zero and one. When the Theil's U-statistics takes the value of zero, it means that the predictive performance of the model is excellant and when it is one, and then it means that the forecasting performance is no better than just using the last actual observation as a forecast.

Table 4. Monte Carlo Experiment: Mean-squared Prediction Errors

	Model 1	**Model 2**	**Model 3**	**Model 4**
Mean	571.60	110.78	43238	14434
Min	469.20	81.93	42507	13958
Max	662.70	181.87	44042	14913
St. Dev.	31.14	13.07	271.91	162.31

Notes: The table shows the summary statistics of MSPEs of various models in Monte Carlo experiment with 500 simulations.

According to Table 4, in repeated samples, Model 2 outperforms Model 1 by a significant margin. The average MSPE for the non-parametric model is 110.78 while this number is 571.6 for the OLS model. We also check the performance of the CW test statistics at the population and show the results in Table 5. The findings confirm our earlier demonstration that, on average, CW test produces a higher statistics for the OLS model while it has worse out-of-sample forecasts than the non-parametric model, and the findings are consistent at each percentile regardless of the selection of the null model. In light of the simulation results, one can conclude that having a statistically significant higher CW test statistics for one model than a second one does not guarantee that the former has better out-of-sample forecasts than the latter.

Table 5. Monte Carlo Experiment: The Clark and West Test Statistics

CW test statistics when the null is driftless random walk											
Percentile	**0.10%**	**0.50%**	**1%**	**2%**	**5%**	**10%**	**50%**	**75%**	**90%**	**95%**	**99%**
Model 1	10.87	10.88	10.89	10.91	10.93	10.95	11.01	11.04	11.07	11.09	11.12
Model 2	9.99	10.02	10.05	10.07	10.10	10.13	10.22	10.25	10.30	10.33	10.38
CW test statistics when the null is random walk with a drift											
Percentile	**0.10%**	**0.50%**	**1%**	**2%**	**5%**	**10%**	**50%**	**75%**	**90%**	**95%**	**99%**
Model 1	10.63	10.65	10.66	10.67	10.70	10.71	10.78	10.81	10.84	10.86	10.90
Model 2	9.38	9.39	9.42	9.45	9.50	9.52	9.64	9.69	9.73	9.76	9.82

Notes: The table shows the Clark and West (CW) test results for Model 1 (OLS) and Model 2 (NP) in a Monte Carlo experiement with 500 replications. The numbers show the test statistics at the 0.1 percentile through 99 percentile. We use rolling regression method with a window of 200 observations. In the NP forecasts, we compute the least-squares cross-validated bandwidths with local linear estimators. The limiting distribution of the CW test under the null is standard normal.

5. Conclusion

The Clark and West test is commonly used in testing the predictive accuracy of alternative exchange rate models, stock return predictability, inflation forecasting and unemployment forecasting. The findings in this paper help better interpreting the Clark and West test statistics and prevent making the wrong conclusion derived from the *magnitude* of the test statistics. Our results show that the Clark and West test fails to reflect the relative performance of a superior model over a relatively weaker model. Even though the MSPE of a superior model is far below a weaker alternative, the test does not reflect this in their test statistics. Hence, as noted in Diebold (2015), one can conclude that the Clark and West test statistics is suitable only to compare a structural model against a random walk null but not suitable for comparison across alternative structural models. The Monte Carlo experiment also confirms this finding. Therefore, practitioners should not put too much emphasis on the *magnitude* of the statistically significant Clark and West tests statistics.

References

Arai, N. (2014). Using forecast evaluation to improve the accuracy of the Greenbook forecast. *International Journal of Forecasting*, *30* (1), 12-19.

Clark, T.E., and West, K.D. (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135, 155–186.

Diebold, F.X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, *33* (1), 1-24.

Diebold, F., and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics,* 13, 253–263.

Dimpfl, T., and Jank, S. (2015). Can internet search queries help to predict stock market volatility?. *European Financial Management*. doi: 10.1111/eufm.12058.

Engel, C., Mark, N.C., and West, K.D. (2015). Factor model forecasts of exchange rates. *Econometric Reviews*, *34* (1-2), 32-55.

Faust, J., and Wright, J.H. (2013). Forecasting inflation. *Handbook of economic forecasting*, *2* (Part A), 3-56.

Ferraro, D., Rogoff, K., and Rossi, B. (2015). Can oil prices forecast exchange rates? An empirical analysis of the relationship between commodity prices and exchange rates. *Journal of International Money and Finance*, *54*, 116-141.

Gregory, A.W., and Zhu, H. (2014). Testing the value of lead information in forecasting monthly changes in employment from the Bureau of Labor Statistics. *Applied Financial Economics*, *24* (7), 505-514.

Hutter, C., and Weber, E. (2015). Constructing a new leading indicator for unemployment from a survey among German employment agencies. *Applied Economics*, *47* (33), 3540-3558.

Löffler, G. (2013). Tower Building and Stock Market Returns. *Journal of Financial Research*, *36* (3), 413-434.

Nikolsko-Rzhevskyy, A., and Prodan, R. (2012). Markov switching and exchange rate predictability. *International Journal of Forecasting*, *28* (2), 353-365.

Sousa, R.M., Vivian, A., and Wohar, M.E. (2016). Predicting asset returns in the BRICS: The role of macroeconomic and fundamental predictors. *International Review of Economics & Finance*, *41*, 122-143.

Theil, H. (1966). *Applied Economic Forecasting*. North-Holland Publishing Co., Amsterdam.

Wang, J., and Wu, J.J. (2012). The Taylor rule and forecast intervals for exchange rates. *Journal of Money, Credit and Banking*, *44* (1), 103-144.

West, K.D. (1996). Asymptotic inference about predictive ability. *Econometrica,* 64, 1067–1084.

Wu, S.J., and Lee, W.M. (2015). Predicting severe simultaneous bear stock markets using macroeconomic variables as leading indicators. *Finance Research Letters*, *13*, 196-204.